Before the
Federal Communications Commission
Washington, DC

In the Matter of	)	
	)	
Protecting the Privacy of Customers of Broadband	)	WC Docket No. 16-106
	)	
and Other Telecommunications Services	)	

Reply comments of Arvind Narayanan and Dillon Reisman

Filed online
June 27, 2016

We are Arvind Narayanan, an assistant professor of Computer Science at Princeton University, and Dillon Reisman, an independent research consultant working with Princeton's Center for Information Technology Policy.

We are responding to comments to the Commission's proposal on new broadband privacy rules that have criticized the FCC for going beyond the FTC's established framework and the White House's Consumer Privacy Bill of Rights, since the rules do not consider the *sensitivity* of data collected or shared.[1] While *sensitive* and *non-sensitive* are important distinctions in general, we believe that for BIAS providers to consider them in protecting customer data would be highly impractical at best and seriously counterproductive at worst. We therefore argue that these considerations should not factor into the FCC's rule making.

Our views arise from three key technical arguments:

1. It is technically infeasible for ISPs to accurately determine which customer data is sensitive.

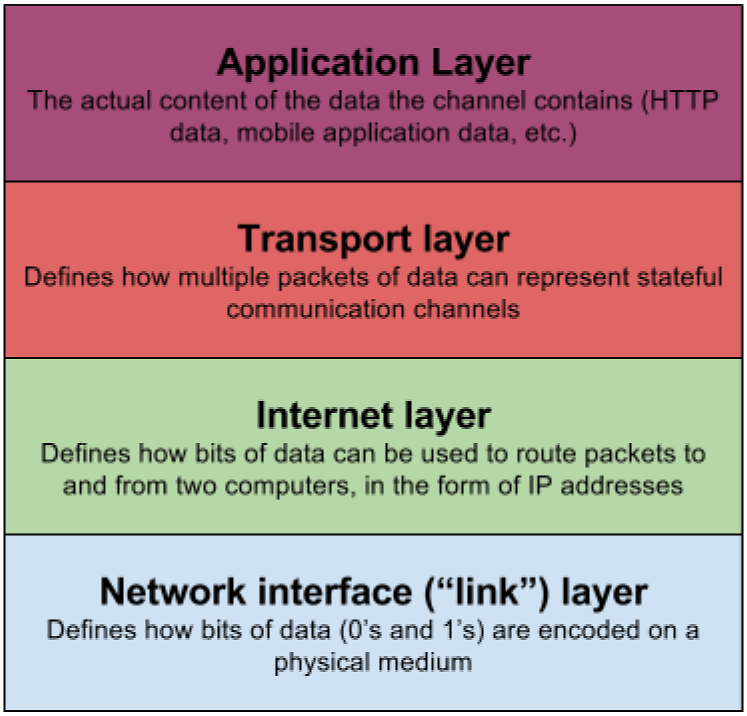
2. Even attempting to make this inference would require significant privacy intrusions, making this rule counterproductive.

3. In the realm of network traffic, the sensitive vs. non-sensitive dichotomy is largely meaningless to begin with, as sensitivity emerges out of the accretion of facts about individuals and the inferences that can be drawn from it.

[1] E.g., Comments of the United States Telecom Association, WC 16-106 ("...the Commission proposes to jettison the long and widely held notion that information varies in its sensitivity and in the context in which it is shared..."); Comments of AAF, et al, WC 16-106 ("...the FCC's proposal inappropriately regulates data based on the type of marketing it is used for, rather than the context or the sensitivity of the data..."); Comments of Commissioner Maureen K. Ohlhausen, FTC, WC 16-106 ("...the FCC would require opt-in consent for many uses of non-sensitive consumer data by BIAS providers, yet would require no consent at all for certain uses of sensitive data by those providers.")

Specifically, we argue that:

I. Determining the sensitivity of data contained in Internet traffic requires ISPs to reverse-engineer application-specific protocols. This is technically infeasible and also creates perverse incentives for privacy violations.



Application Layer
The actual content of the data the channel contains (HTTP data, mobile application data, etc.)

Transport layer
Defines how multiple packets of data can represent stateful communication channels

Internet layer
Defines how bits of data can be used to route packets to and from two computers, in the form of IP addresses

Network interface ("link") layer
Defines how bits of data (0's and 1's) are encoded on a physical medium

The TCP/IP networking model. The Internet's architecture calls for each layer to be agnostic to the details of the layers above it.[2]

Networks are designed so that each layer is agnostic to the details of the layers above it. In particular, ISPs are agnostic to the details of applications and content. This engineering principle is closely tied to the principle of net neutrality. The key difficulty, then, is this: the "semantics" or meaning of a piece of content is specific to each application, of which there are

[2] Microsoft TechNet, "The TCP/IP Model," accessed June 27, 2016 (https://technet.microsoft.com/en-us/library/cc786900(v=ws.10).aspx)

thousands if not millions.[3] The Internet is engineered in such a way that ISPs need not — and generally, cannot — infer the semantics of applications, that is, whether a particular piece of data represents (say) a customer's social security number, health information, a video, or something else. Applications define their own semantics, and there is no central repository of application semantics.

Let us examine the technical steps necessary for an ISP to infer anything about the sensitivity of content. For a start, it would require Deep Packet Inspection, which already goes beyond what is technically required for an ISP to perform its basic function of routing Internet traffic — content inspection could be considered a privacy violation in and of itself.[4] We agree that requiring sensitive data classification would create a "perverse incentive for BIAS providers to identify or inspect protected data" as others have suggested.[5] But it would take far more: the ISP would have to decode and reverse-engineer each application-specific communication protocol that it encounters, determine which data is sensitive, and apply that characterization in an automated fashion to all future data collected from that application. Doing this for even a single proprietary application would be a research and engineering challenge, and doing it for all traffic is far beyond the realm of possibility.

[3] The "application layer" refers to protocols such as HTTP. Applications in the more common sense of the term can be thought of as residing in a further layer above the "application layer."

[4] Daly, The Legality of Deep Packet Inspection (2010), (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1628024)

[5] Comments of the Electronic Frontier Foundation, WC 16-106 at 5

II. IP addresses of web traffic alone can reveal sensitive information about a customer, while not being strictly categorizable as sensitive data.

To further illustrate why it is difficult for ISPs to categorize data as sensitive or non-sensitive, let us consider IP addresses. IP addresses can often be uniquely mapped to a domain name, i.e., a website. In some cases, knowing that a customer visited a website will not reveal sensitive information. A customer who sends traffic to the IP address 216.58.194.46, for instance, will have only revealed to their ISP that they have visited the domain www.google.com, without sharing the content of their search.

In many other cases, however, a domain name alone can reveal a multitude of sensitive attributes about a customer. If a customer visits domains related to their health, like www.mesothelioma.com or www.breastcancer.org, then the IP addresses will reveal that the customer has an interest in those particular health problems. If a customer visits domains related to their finance, like www.collegeloan.com or www.mortgage101.com, then the IP addresses will reveal finance-related attributes about the customer.

IP addresses, which are regularly sampled and retained by ISPs,[6] could represent any number of sensitive domains that a customer visits on the web. Only through a deeper analysis of each IP address visited could an ISP know that they had inadvertently collected a sensitive webpage visit from their customer — it would be infeasible to assign a blanket categorization to IP addresses in general.

Similarly, encrypted traffic defies simple categorization, because even if the ISP cannot decrypt the traffic, it may be possible to infer some sensitive attributes. In particular, Swire and

[6] Comment of Nick Feamster, WC 16-106 at 3

others have claimed that when web traffic is encrypted, an ISP cannot read the full URL of a website that a customer visits.[7] Computer science research, however, has shown that this assumption does not always hold. Researchers at UC Berkeley have found that encrypted web traffic can be used to infer the pages within an encrypted site that a customer visits.[8] A study from the University of Cambridge showed that the amount of data transmitted over encrypted connections could also be used to infer the pages a customer visits.[9] The knowledge of full URLs visited by a customer can be even more revealing than IP addresses. For example, though a visit to the Wikipedia homepage ("https://en.wikipedia.org") does not reveal any interesting user information, an actor who could infer that an encrypted page visit was made to "https://en.wikipedia.org/wiki/Mesothelioma" would learn a lot about a user.

III. A collection of seemingly non-sensitive data can be used to infer sensitive attributes

Machine learning can be employed to infer sensitive information out of seemingly non-sensitive attributes. Researchers have shown that relatively simple machine-learning algorithms can be used to infer sensitive personal attributes from users' "Likes" on Facebook — these attributes include sexual orientation, ethnicity, religious and political views, use of addictive substances, age, and gender.[10] Phone metadata, often considered less sensitive than

[7] Swire, et al., Online Privacy and ISPs: ISP Access to Consumer Data is Limited and Often Less than Access by Others (2016) http://www.iisp.gatech.edu/sites/default/files/images/online_privacy_and_isps.pdf; For example, for the full url http://www.wired.com/**2016/06/remarkable-tech-bringing-deaf-hearing-worlds-together/,** the bolded section will not be visible in encrypted web traffic.

[8] Miller, et al., I Know Why You Went to the Clinic: Risks and Realization of HTTPS Traffic Analysis (2014) (https://www.petsymposium.org/2014/papers/Miller.pdf)

[9] Danezis, Traffic Analysis of the HTTP Protocol over TLS (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.3893&rep=rep1&type=pdf)

[10] Kosinski, et al., Private traits and attributes are predictable from digital records of human behavior (2013) (http://www.pnas.org/content/110/15/5802.abstract)

the contents of conversations, can be used to infer sensitive information about individuals such as medical condition or firearm ownership.[11]

These studies succeed because many non-sensitive behaviors correlate to a sensitive characteristic. An individual who "Likes" a certain TV show or news channel might be more likely to belong to a particular political party. It has proved futile enumerate all attributes that might correlate with a sensitive characteristic. Machine-learning techniques are capable of utilizing hundreds or thousands of seemingly innocuous observations about an individual to construct an accurate and revealing profile. Any type of attribute or behavior can useful in this process, making the sensitive vs. non-sensitive dichotomy essentially meaningless.

Conclusion

In our reply, we have shown that attempting to use the distinction between sensitive and non-sensitive data in rulemaking is (a) conceptually infeasible, since many types of attributes defy simple categorization (b) scientifically infeasible, since advanced techniques can be used to infer sensitive attributes out of supposedly "non-sensitive" data, and (c) infeasible in terms of engineering, since ISPs have no known way to identify sensitive data in Internet traffic, even if we can agree on what constitutes sensitive data. While some commenters admirably point to the leadership of the FTC and the White House Consumer Privacy Bill of Rights in using sensitive data as a guide in rulemaking, in the context of broadband Internet providers the sensitive data

[11] Mayer and Mutchler, Metaphone: The Sensitivity of Telephone Metadata (2014) (http://webpolicy.org/2014/03/12/metaphone-the-sensitivity-of-telephone-metadata/)

distinction does not make sense. We believe the Commission is right to use a different standard in its rulemaking.